Interpreting the results of observational research: chance is not such a fine thing

Paul Brennan, Peter Croft

In a randomised controlled trial, if the design is not flawed, different outcomes in the study groups must be due to the intervention itself or to chance imbalances between the groups. Because of this tests of statistical significance are used to assess the validity of results from randomised studies. Most published papers in medical research, however, describe observational studies which do not include randomised intervention. This paper argues that the continuing application of tests of significance to such non-randomised investigations is inappropriate. It draws a distinction between bias and chance imbalance on the one hand (both randomised and observational studies can be affected) and confounding on the other (a unique problem for observational investigations). It concludes that neither the P value nor the 95% confidence interval should be used as evidence for the validity of an observational result.

Epidemiologists and clinical researchers design studies to estimate the effect which a presumed cause or treatment has on the occurrence of a disease. Most questions about causes of disease cannot be addressed by experiments: we must rely on the observation of life as it is, rather than of the results of controlled intervention. Such observational studies cannot provide proof of causality but are still the basis for reasoned public health decisions.

In the presentation of results from observational studies significance tests are often presented as judgments on the "truth" or validity of the effect which a presumed cause has on the occurrence of a disease. In 1965 Bradford Hill lamented this application of statistics,[1] a concern given prominence again recently.[2] Yet almost 30 years on, phrases such as "the result just failed to reach statistical significance" are still part of the argot of medical papers and presentations. The move towards estimating confidence intervals has not resolved this problem, as the 95% confidence interval is often presented as if it was a significance test—"the confidence interval does not quite include one, so the relative risk is statistically significant."

The question we have addressed is whether statements of statistical probability should ever be used to judge the validity of findings from observational studies. This question has been argued over for many decades, notably in psychological publications[3] and more recently in epidemiological journals.[4] In this paper we review the concepts of bias, confounding, and statistical tests of probability as they apply to the interpretation of experimental studies and then consider them in the context of observational studies.

ARC Epidemiological Research Unit, University of Manchester Medical School, Manchester M13 9PT
Paul Brennan, *statistician*
Peter Croft, *epidemiologist*

Correspondence to:
Mr Brennan.

BMJ 1994;309:727-30

BIAS AND CONFOUNDING

The concepts of bias and confounding are central to the ideas reviewed in this paper. To illustrate their meaning we take the example of oral contraceptive use and cervical cancer and assume that in truth the pill does not cause this cancer.

Consider a study in a population of women in whom, as expected, the actual rate of cervical cancer in those exposed to the oral contraceptive is identical with the rate in unexposed women (rate ratio = 1). A fault in the study design leads to a misrepresentation of these rates and the appearance of a link between exposure and disease (for example, an observed rate ratio of 2). This is bias. Though there are many possible sources of bias in a study (Sackett noted 35 separate forms),[5] they usually originate either in the selection of subjects for a study or in the information collected from study participants. As an example of the latter, women taking oral contraceptives may be more likely to have regular cervical smear tests than women not taking the pill. As a result, women who have not used the pill may seem to have a reduced rate of disease because of the lower opportunity for diagnosis.

Definitions of bias and confounding that may affect interpretation of study results

- *Bias:* Faults in study design lead to a misrepresentation of the relation between exposure and disease
- *Confounding:* Exposure seems to be associated with disease. However, the relation exists only because the exposure is associated with other risk factors for the disease and not because the exposure causes the disease

Next consider a similar study in a different population of women in whom pill users differ from non-users in other ways which do cause cervical cancer—their rates of genital infection, for example. Because of this the rates of disease among pill users and non-users in this population are actually different (rate ratio = 2·0). The study design is adequate and the difference is reflected in the study result. No bias is present. However, though the result indicates that the exposure is associated with the outcome, it does not mean that the exposure has caused the outcome. Pill users in this population were more likely to acquire genital infections than non-users. As such infections are associated with cervical cancer, disease rates will be higher in the users. Pill use is linked with cervical cancer because it is associated with infection. This is confounding.

Bias is an issue of study design whereas confounding is an issue of alternative explanations of the study result. It should be noted that a study result may be subject to both bias and confounding. A third influence on a study result is that of chance. The relation between chance, bias, and confounding forms the topic of this paper.

Randomised controlled trials

A randomised controlled trial is an attempt to assemble two groups of subjects who are similar in all respects, apart from the intervention under investigation. The rationale is that randomisation ensures that the allocation of treatment is independent of other exposures which may affect outcome. The randomised trial can therefore uniquely rule out the possibility of confounding as an explanation of the result. This should mean that any difference in outcome between the two groups is attributable to the intervention. However, two alternative explanations must be considered—(*a*) that the difference is due to bias, and (*b*) that randomisation has, by chance, resulted in two groups which are not comparable. To illustrate this we used the example of a trial of folic acid to prevent neural tube defects.[6]

Women with a previous pregnancy complicated by a neural tube defect and who were planning another pregnancy were randomised to receive either folic acid or a placebo. A total of 910 women received folic acid and 907 placebo. The proportion of women in each group whose subsequent pregnancy had a neural tube problem is shown in table I. Altogether 1195 women became pregnant during the study, 27 of the pregnancies being complicated by a neural tube defect. Six of these were in the group randomised to take folic acid and 21 in women who received placebo. Folic acid appeared to prevent neural tube defects, women in the placebo group having more than a threefold risk compared with the actively treated group.

TABLE I—*Numbers of pregnancies complicated by neural tube defect according to whether women were randomised to receive folic acid or placebo. Randomised controlled trial*

	Intervention	
	Folic acid	Placebo
Neural tube defects	6	21
No of pregnancies	593	602
Risk of affected pregnancy	1·0%	3·5%
Risk ratio for placebo:folic acid	3·5/1·0=3·5	

BIAS

Bias could explain this observation. For example, if patients had known their treatment status those taking folic acid might have chosen to supplement their diets in other ways. Minimising bias is an issue of study design. However, neither patients nor medical staff were aware of the participants' treatment group—that is, the trial was double blind.

CHANCE IMBALANCE

Another explanation might be that the observed results were simply due to chance. If folic acid was no different from placebo and the study was unbiased, then the observations in table I could have arisen only if randomisation had resulted in groups which were not comparable—that is, if higher risk women had been allocated by chance to the placebo group. This phenomenon, in which randomisation results in an unequal distribution of risk factors, has been termed random confounding.[4] As it is simply an imbalance in the treatment groups arising by chance we have adopted the term "chance imbalance." A basic strength of clinical trials is that they can be made as large as necessary to ensure that imbalances in randomisation are extremely unlikely (in contrast with bias, which will not diminish as sample size increases). Also it is possible to calculate the probability of any observed difference occurring by chance when, in fact, no real difference exists between the groups. This probability is the P value.

When calculating the P value for the data in table I we assume that if folic acid and placebo had identical effects 27 women would have given birth to a child with a neural tube defect regardless of the treatment—with roughly half in each group. The P value tells us how likely it is that as few as six would end up in one group and 21 or more in the other group given this assumption of identical effectiveness. The answer ($P=0·003$) informs us that, assuming there is no real difference between placebo and active treatment, the probability of such an uneven randomisation would be about one in 300. This suggests that chance imbalance is an unlikely explanation for the finding.

The P value approach assumes the true difference between the exposures to be zero and evaluates the probability that the observed effect is due to chance imbalance. The confidence interval approach shifts attention to the effect actually observed in the study and calculates a region around it where the true effect is likely to be. The true effect can lie anywhere and so the size of this region is restricted by how certain we wish to be that it does encompass the true effect. This is done by assuming that an extreme example of chance imbalance has not occurred and that the true effect is not very far from the observed effect. The definition of extreme is arbitrary, but the usual method is to exclude the 5% most extreme possibilities of chance imbalance. This assumption underlies the 95% confidence interval. It would, however, be equally reasonable to exclude the 1% most extreme possibilities for a 99% confidence interval or the 10% most extreme for a 90% confidence interval. In our example the observed risk ratio of 3·5 for a subsequent fetal defect among women not taking vitamin supplementation has a 95% confidence interval of 1·4 to 8·5. Only if an extreme example of chance imbalance has occurred will the true effect lie outside the region.

This evidence against chance imbalance as an explanation of the effect of folic acid, together with the precautions taken to exclude bias, led the study group to conclude: "The result is unlikely to be due to chance and the randomised double blind design excludes bias as an explanation."[6] The only remaining hypothesis is that vitamin supplements taken by women at high risk caused a reduction in the number of neural tube defects complicating a subsequent pregnancy.

Observational studies

Much epidemiological research is not experimental but entails observation of what occurs without controlled intervention. The paradigm of observational research is the cohort study, though the following arguments apply also to case-control studies. Superficially there are many similarities between a clinical trial and a cohort study. Both usually include two groups of subjects being followed up over time with the effect of an exposure or intervention as the main focus of interest. Bias and chance imbalance are also relevant to cohort studies. However, there is a crucial difference between the two types of study. In a clinical trial the exposure or intervention status of each subject is decided at the start of the study by randomisation. By contrast in a cohort study each subject chooses or has arrived at this status before the study. This difference is fundamental to the interpretation of results obtained from cohort studies and has important implications for the meaning of any statistics computed.

CONFOUNDING

In addition to bias and chance imbalance, confounding must be considered in a cohort study. This arises when there are differences between subjects who have "chosen" to be exposed and those who have not, with these differences being separately related to the disease under study.

In a randomised clinical trial there can be no confounding, and assessment of chance imbalance has meaning once the possibility of bias has been removed. However, a similar assessment in cohort studies requires the absence of both bias and confounding. If confounders are present, then an effect may be observed which is accompanied by small P values and tight confidence intervals but which does not represent a causal effect and cannot be explained by chance or bias. The assumption of causality based on such results is flawed. Attempts must therefore be made to remove the effect of confounders before assessing the extent of chance imbalance. It is impossible, however, to be sure that all confounders have been identified in a cohort study.[7] Moreover, whereas a clinical trial may be made sufficiently large to reduce the possibility of chance imbalance to a predetermined level, the effects of confounders in observational studies do not diminish with increasing sample size. As a consequence, if a confounder is not recognised and adjustments made for its effect the exposed and unexposed groups in such studies will not be comparable.

CHANCE IMBALANCE

The consequence is that P values in an observational study may not represent the probability (under the null hypothesis of no real difference between the groups) that chance imbalance is the cause of any observed differences. Similarly, confidence intervals cannot be defined as regions with a certain probability of containing the true effect. Why then should they be used at all in the interpretation of observational studies? One justification is that their total exclusion would mean the same emphasis being placed on the findings of a small cohort study as on those from a large one. If chance imbalance is contributing to the total distortion of a study result its potential effect will be greater the smaller the study. As an indication of the possible influence of chance imbalance the confidence interval has a part to play in the presentation of the results of observational studies. Yet the confidence interval cannot be interpreted as containing the true effect measure with a certain probability because we cannot know the extent of confounding. The P value has no direct interpretaion and conveys even less information in an observational study than the confidence interval.

In order to illustrate this we use another example of vitamin supplements in the prevention of neural tube defects, this time from a cohort study.[8] The incidence of pregnancies complicated by a neural tube defect in 438 women who had had a previous pregnancy resulting in such an outcome was determined. A total of 178 of the women had taken periconceptional vitamin supplements and they were compared with 260 unsupplemented or control women. The authors found an eightfold increased risk for women who had not supplemented their diets with vitamins (table II).

The investigators thought that this effect was unlikely to be due to bias between the groups. Subsequent correspondence, however, pointed to the possibility of selection bias. The supplemented sample was more likely to come from geographical areas with a low incidence of neural tube defects compared with the control group.[9] However, even in the absence of bias the authors felt unable to interpret the 95% confidence interval for the effect (which ranged from 1·2 to 67·0) as representing the region where the true effect should lie with a probability of 95%. This was because they could not rule out the explanation that something other than vitamin supplementation reduced the incidence of neural tube defects in the pregnancies of the supplemented group of women—that is, they could not exclude confounding. For instance, women who choose to supplement their diets with vitamins periconceptionally may also reduce their smoking and alcohol intake and alter their diet in other ways which affect risk.

TABLE II—*Numbers of pregnancies complicated by neural tube defect according to whether women chose to take vitamin supplements. Cohort study*

	Vitamin supplementation	
	Yes	No
Neural tube defects	1	13
No of pregnancies	178	260
Risk of affected pregnancy	0·6%	5·0%
Risk ratio	5·0/0·6=8·3	

Given that the authors could not allow for such alternative explanations, they admitted that the possibility of confounding presents "an almost untestable hypothesis." What they had was an observed effect with no way of proving whether it represented a causal relation or a byproduct of a confounder.

The inadequacy of using confidence intervals in observational studies was further highlighted three years later when the same investigators released the results from a second cohort of women and combined them with the first.[10] The combined results showed a total of three neural tube defects in the pregnancies of 429 women who received vitamin supplements and 24 among 510 women who had not been supplemented, representing a risk ratio of 6·7 and a 95% confidence interval of 2·0 to 22·2. Clearly the confidence interval has narrowed because of the larger numbers. However, if confounders exist, then the risk ratio from the combined cohort remains just as confounded as that from the first cohort. So the tighter confidence interval remains centred on the confounded effect estimate and not on the true effect estimate. Indeed, correspondence again pointed to possible differences between the two groups: the proportion of women in social classes I and II was twice as high in the supplemented groups as in the unsupplemented.[11] Paradoxically, in the presence of confounding, as a study gets larger the probability that the true effect estimate lies within the limits of a 95% confidence interval decreases. Consequently, a consideration of confounding is of paramount importance before any consideration of the role of chance.

Bradford Hill revisited

Bradford Hill emphasised the need to weigh critically the evidence or lack of it for alternative explanations of study results and to renounce the glitter of the *t* test.[1] Greenland in 1987 observed that Bradford Hill's comments on the misuse of significance testing were the most neglected portion of this paper.[12] Over the past decade there has been much discussion of the advantages of confidence intervals over P values, which has shifted the emphasis of inference away from decision making (the original role of P values) towards assessing the size and precision of observed effects.[13-15]

In addition to voicing concern about the use of statistical inference, Bradford Hill discussed the problems of inferring causality from observational associations.[1] His paper has been used frequently as a checklist of "criteria" for causality and because of this it has come under attack.[16] In fact, Bradford Hill did not use the word "criteria" once in the paper, and it is clear that he never intended such a checklist.[12] His central concern was the need to make considered assessments of possible alternatives to causality when explaining observed associations. He relegated chance to a minor position in this process.

The search for confounding is all about the search for alternative explanations. Randomised controlled trials, which uniquely can tackle confounders which are unknown or unmeasured, give us the most powerful evidence relating to cause and effect. However, most public health questions cannot be addressed through randomised controlled trials and must rely on the results of observational studies. The problem of unknown confounders means that observational studies cannot provide proof of causality, and statistical tests cannot help. Yet providing proof has never been the function of epidemiology, and decision making in public health has not been paralysed as a result. Instead what is required is "the search for any other way of explaining the set of facts before us."[1] It is time for writers, presenters, reviewers, and editors to grasp the nettle and put Bradford Hill into practice.

Summary points

- In presenting and discussing the results of an observational study the greatest emphasis should be placed on bias and confounding
- Confidence intervals should be relegated to a small part of both the results and discussion sections as an indication, but no more, of the possible influence of chance imbalance on the result
- 90% Confidence intervals should be accepted as readily as 95% confidence intervals
- The term "statistical significance" should not be used and P values should not be published in observational studies

Recommendations

On the basis of these arguments we make the following practical suggestions regarding the use of probability statistics in observational studies.

(1) In presenting and discussing the results of an observational study the greatest emphasis should be placed, firstly, on questions of bias within the study and, secondly, on the issue of confounding.

(2) Confidence intervals should be relegated to a small part of both the results and discussion sections as an indication, but no more, of the possible influence of chance imbalance on the result. The move away from using confidence intervals as a surrogate for statistical significance would be helped if 90% confidence intervals were accepted as readily as 95% confidence intervals.

(3) The structure of any discussion would then be in terms of: (*a*) to what extent might flaws in the study design have biased the study result?; (*b*) if the result is thought to be free of bias, to what extent might other causes have confounded the observed association?; and (*c*) if the result appears unbiased and known confounders have been accounted for, what might have been the extent of chance imbalance as represented by the confidence interval?

(4) The term "statistical significance" should not be used and P values should not be published in observational studies.

We thank Drs P Hannaford, S Khan, A Silman, D Symmons, the statistical referee, and Dick Keystone for helpful comments.

1 Hill AB. The environment and disease: association or causation? *J R Soc Med* 1965;**58**:295-300.

2 Jolley D. The glitter of the t table. *Lancet* 1993;**342**:27-9.

3 Morrison DE, Henkel RE. *The significance test controversy*. London: Butterworths, 1970.

4 Greenland S. Randomization, statistics, and causal inference. *Epidemiology* 1990;**1**:421-9.

5 Sackett DL. Bias in analytic research. *Journal of Chronic Diseases* 1979;**32**: 51-63.

6 MRC Vitamin Study Research Group. Prevention of neural tube defects: results of the MRC vitamin study. *Lancet* 1991;**338**:131-7.

7 Davey-Smith G, Phillips AN, Neaton JD. Smoking as an 'independent' risk factor for suicide: illustration of an artefact from observational epidemiology? *Lancet* 1992;**340**:709-12.

8 Smithells RW, Shephard S, Schorah CJ, Seller MJ, Nevin NC, Harris R, *et al.* Possible prevention of neural tube defects by periconceptional vitamin supplementation. *Lancet* 1980;i:647.

9 Stone DH. Possible prevention of neural tube defects by periconceptional vitamin supplementation. *Lancet* 1980;i:647.

10 Smithells W, Seller MJ, Harris R, Fielding DW, Schorah CJ, Nevin NC, *et al.* Further experience of vitamin supplementation for prevention of neural tube defect recurrences. *Lancet* 1983;i:1027-31.

11 Rose G, Cooke ID, Polandi PE, Wald NJ. Vitamin supplementation for prevention of neural tube defect recurrences. *Lancet* 1983;i:1165-6.

12 Greenland S. *Evolution of epidemiological ideas*. Chestnut Hill, Massachusetts: Epidemiology Resources, 1987.

13 Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *BMJ* 1986;**292**:746-50.

14 Rothman K. A show of confidence. *N Engl J Med* 1978;**299**:1362-3.

15 Poole C. Beyond the confidence interval. *Am J Public Health* 1987;**77**:195-9.

16 Lanes SF. The logic of causal inference. In: Rothman KJ, ed. *Causal inference*. Chestnut Hill, Massachusetts: Epidemiology Resources, 1988:59-75.

(Accepted 6 June 1994)

Some observations on attempts to measure appropriateness of care

Nicholas R Hicks

Department of Public Health and Health Policy, Oxfordshire Health Authority, Headington, Oxford OX3 9DZ
Nicholas R Hicks, *consultant public health physician*

BMJ 1994;309:730-3

There are a growing number of published studies that suggest that much health care is delivered inappropriately. There are calls for measures of appropriateness to be used by purchasers and others to regulate or influence the delivery of health care. This paper explores assumptions inherent in results generated by a leading measure of appropriateness and concludes that there are considerable uncertainties about the measure's meaning, the magnitude of bias that it contains, and the degree to which its application can be generalised. Some of these uncertainties could be resolved if the tacit assumptions inherent in the generation of the criteria could be made explicit. Existing measures of appropriateness are not yet sufficiently robust to be used with confidence to influence or control the delivery of health care. They may have a use as an aid rather than as a constraint in clinical decision making. A randomised controlled trial could resolve whether patients achieve better outcomes if their care is influenced by appropriateness criteria.

A leading article by Brook published in the *BMJ* recently identified appropriateness as "the next frontier" in the development of clinical practice.[1] It argued that, firstly, there is too much information about medical practice for any doctor to assimilate all the information relevant to their practice. It is therefore impossible to "practise good medicine without additional help." Secondly, for this (and other reasons) many patients receive care that is "inappropriate" (contributing to overuse of health care) and many others are not offered "appropriate" care (underuse of health care). Thirdly, the appropriateness of care can be measured, and, finally, the application of measures of appropriateness can reduce or eliminate both overuse and underuse of medical interventions.

These claims, if true, have huge implications for medical practice, given that some studies estimate that 20% to 60% of care is less than appropriate.[2] These findings have led to calls for the profession, patients, and purchasers of care to use measures of appropriateness to regulate the delivery of care. Before such measures are used to influence care in the United Kingdom, it seems reasonable to explore the meaning of appropriateness scores to ask if the results could be biased and to consider how well judgments about appropriateness can be generalised. It is even pertinent to ask what the phrase "appropriate care" means. Brook and colleagues at the Rand Corporation and the University of California at Los Angeles (UCLA) have developed and pioneered the use of one of the leading tools for measuring appropriateness of care.[3] I explored the question of whether measures of appropriateness are sufficiently robust to apply to everyday practice in the United Kingdom by examining the process by which Rand appropriateness scores are generated.